Université de Caen Normandie
IUT Grand Ouest Normandie – Pôle de Caen
Département Science des Données

Diplôme Bachelor Universitaire de Technologie
SCIENCE DES DONNÉES
Parcours Science des Données : Exploration et modélisation statistique

Deuxième année

PORTFOLIO

Lou-Anne THOMAS

Année universitaire 2023-2024

Table des matières

SAÉ 3.01 - Recueil et analyse de données par échantillonnage ou plan d'expérience....3
SAÉ 3.02 – Intégration de données dans un datawarehouse....4
SAÉ 3.03 – Description et prévision de données temporelles....5
SAÉ 3.04 – Conformité réglementaire pour analyser des données....6
SAÉ 4.01 – Expliquer ou prédire une variable quantitative à partir de plusieurs facteurs....7
SAÉ 4.02 – Reporting d'une analyse multivariée....8

SAÉ 3.01 – Recueil et analyse de données par échantillonnage ou plan d'expérience

A. Présentation

J'ai effectué ce travail seul sur un créneau horaire défini, totalisant 4 heures, au cours desquelles j'ai mis en place un plan d'expérience à l'aide du logiciel **RStudio**. L'objectif de ce projet était de déterminer la composition optimale des ingrédients d'une pâte à pizza afin de maximiser sa qualité gustative. Pour ce faire, nous disposions de scores attribués à chaque recette variant selon les quantités de farine, de sel et de levure. L'analyse visait donc à déterminer les proportions idéales de chaque ingrédient influençant le goût de la pizza.

B. Démarche et résolution du problème

La démarche de l'étude a impliqué la définition de niveaux haut et bas pour les facteurs clefs de la pizza (farine, sel, levure). Pour chaque recette, deux répliques de la pâte ont été cuisinées. La figure ci-contre (cf. Figure 1) représente la matrice de conception où le niveau haut (+) représente la teneur élevée de l'ingrédient tandis que le niveau bas (-) représente la teneur faible de l'ingrédient. On y retrouve la recette (Ord), la réplique de la recette (1 ou 2) et le score attribué à la recette. De plus, nous avons 8 configurations expérimentales. Ces configurations correspondent aux différentes combinaisons des niveaux haut et bas des ingrédients (farine, sel, levure) dans chaque recette, permettant ainsi d'évaluer l'effet de chaque ingrédient sur la qualité gustative de la pizza.

repl	flour	salt	bakPow	Score	Ord
1	-	-	-	5.33	7
1	+	-	-	6.99	1
1	-	+	-	4.23	4
1	+	+	-	6.61	8
1	-	-	+	2.26	5
1	+	-	+	5.75	2
1	-	+	+	3.26	6
1	+	+	+	6.24	3
2	-	-	-	5.70	7
2	+	-	-	7.71	1
2	-	+	-	5.13	4
2	+	+	-	6.76	8

Figure 1 – Matrice de conception

Après avoir dressé cette matrice, j'ai pu mettre en place mon modèle de régression de type ANOVA en prenant en compte les interactions entre les facteurs. D'après ce modèle et l'étude des coefficients, il a été possible d'identifier les facteurs significatifs, en l'occurrence la farine et la levure. Un modèle simplifié, sélectionnant les facteurs significatifs (excluant le facteur levure et les interactions), a ensuite été développé, mettant en évidence l'influence de ces deux facteurs sur le score gustatif. J'ai pu, à l'aide de mon modèle simplifié, prédire des scores pour chacune de mes conditions expérimentales (les 8 configurations possibles).

Ainsi, mes prédictions ont mis en évidence qu'une teneur élevée en farine influence positivement le score attribué à la pâte et, à l'inverse, une teneur élevée en levure influence ce score négativement. Le graphique ci-contre (cf. Figure 2) permet de visualiser ces impacts. En effet, l'on observe que si la teneur en farine (en bleu) passe du niveau bas au niveau haut, alors le score obtenu augmente de 1,23 point. À l'inverse, si la teneur en levure (en rouge) passe du niveau bas au niveau haut, alors le score diminue de 1,87 point.

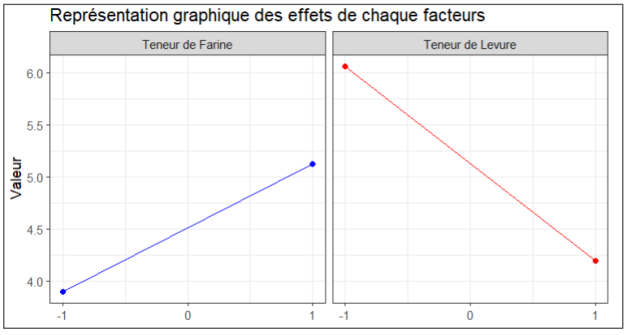

Figure 2 – Représentation des effets de la farine et de la levure

C. Conclusion

Ce projet m'a permis d'étudier des modèles statistiques plus complexes en prenant en compte simultanément plusieurs facteurs, et en examinant les effets de ces facteurs ainsi que leurs interactions. La mise en place de ce plan d'expérience a permis de réduire le nombre d'essais nécessaires. En effet, sans plan d'expériences, l'exploration des différentes combinaisons d'ingrédients aurait été laborieuse, nécessitant davantage de temps, d'argent et de matériaux.

SAÉ 3.02 – Intégration de données dans un datawarehouse

A. Présentation

Ce projet avait pour objectif de nous confronter aux problématiques liées à l'intégration de données hétérogènes. Lors de ce travail, il a été nécessaire d'intégrer les données de 37 fichiers CSV dans un entrepôt de données en **MySQL**. L'ensemble des données recueillies sur le site *Airvivo* portaient sur des mesures d'indicateurs de la qualité de l'air (concentration d'azote, de monoxyde de carbone, etc.) et météorologiques (vitesse du vent, radiation, etc.) concernant la ville de Sheffield entre 2000 et 2022. Ce travail de groupe (3 personnes) a donné lieu à l'élaboration d'un cahier des charges concernant l'intégration des données dans un entrepôt, d'un rapport d'analyse sur la qualité de l'air de la ville ainsi que d'une soutenance orale permettant de restituer nos résultats.

B. Démarche et résolution du problème

En vue de réaliser une analyse, nous avons suivi les différentes étapes énoncées ci-dessous : conception de l'entrepôt de données, traitement et chargement des données et enfin, l'analyse.

Conception de l'entrepôt de données : Dans cette première partie du projet, nous avons réfléchi à la structure de notre base de données et identifié les ajustements requis pour la transformation de nos données. Pour cela, nous avons mis en œuvre un cahier des charges. Nous y avons défini la structure de l'entrepôt en présentant : un schéma conceptuel et relationnel, la structure des différentes tables, la définition des métadonnées, les contraintes (clés primaires et étrangères, conventions de codage, etc.).

Traitement et chargement : Pour intégrer les données dans notre entrepôt, nous avons automatisé la lecture et le traitement (transformation des données, normalisation des données, insertion au sein de tables, etc.) de ces fichiers au moyen du langage de programmation **Python** et de la bibliothèque **Pandas**. Pour cela, nous avons mis en place un notebook Jupyter documentant le code. Les données ont finalement été insérées dans notre base de données **MySQL** à l'aide de la librairie **Mysql.connector** de **Python**.

Analyse : Pour faciliter nos analyses sous **Python**, nous avons établi des vues SQL permettant de regrouper et d'agréger des données de différentes tables au sein d'une même vue. Par exemple, nous avons créé une vue « Indicateur » qui regroupe les mesures de concentration de plusieurs molécules. Cette vue nous a ensuite permis de déterminer la valeur maximale mesurée chaque jour pour chacune des molécules (cf. Figure 3). Nous avons par la suite réalisé des représentations graphiques à l'aide des librairies **Matplotlib**, **Numpy** et **SciPy** de **Python**.

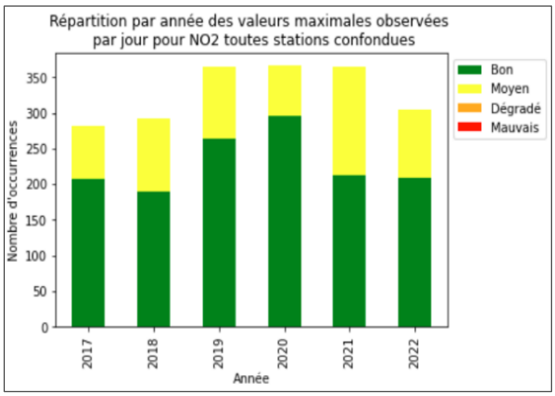

Figure 3 - Extrait d'une représentation graphique

C. Conclusion

Ce projet nous a permis de concevoir notre propre base de données et de découvrir l'existence des vues SQL. Le travail de réflexion et la mise en place du cahier des charges ont été cruciaux, permettant de clarifier nos besoins et attentes. Nous avons également soulevé une problématique essentielle dans la mise en œuvre de ce travail : l'importance des tests. En effet, après avoir constaté l'arrondi non désiré de certaines valeurs au sein de la base de données, nous avons été obligés de procéder à une seconde étape d'insertion des données. Des tests préalables sur quelques lignes auraient évité de perdre un temps considérable à réinsérer des millions de lignes dans la base de données.

SAÉ 3.03 – Description et prévision de données temporelles

A. Présentation

Lors de ce projet, nous avons examiné une série chronologique analysant le nombre mensuel de nuitées enregistrées dans les hôtels de Normandie sur la période de janvier 2011 à septembre 2023. Notre objectif était de créer un modèle prédictif pour l'année 2024 du nombre de nuitées attendu dans les hôtels de Normandie. Pour cela, nous avons travaillé en groupe de 3 étudiants sur les données de l'INSEE à l'aide du langage de programmation statistique **R**. Ce travail a abouti à la rédaction d'un rapport d'analyse suivi d'une présentation orale.

B. Démarche et résolution du problème

Le chronogramme ci-contre permet de visualiser l'entièreté de la série chronologique avec, en rouge, la prévision réalisée correspondant au nombre de nuitées attendues pour l'année 2024. On retrouve sur l'axe des abscisses les mois de chaque année depuis 2011 et, sur l'axe des ordonnées le nombre de nuitées. Pour parvenir à un tel résultat, nous avons suivi les différentes étapes présentées ci-dessous.

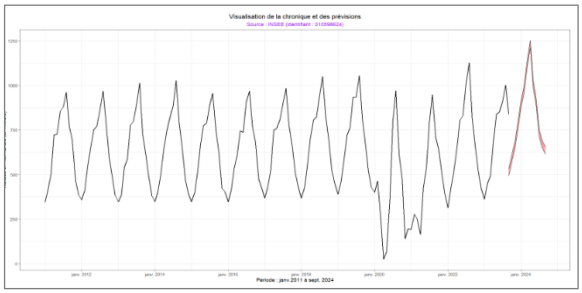

Figure 4 - Chronogramme de la série temporelle

Analyse exploratoire : Dans l'objectif de mieux comprendre nos données, nous avons visualisé la série en utilisant des visualisations telles qu'un chronogramme pour détecter les tendances et les variations saisonnières. La tendance représente l'évolution générale des observations sur une période donnée, tandis que les saisons correspondent à des schémas répétitifs. Ainsi, si l'on se réfère à la Figure 4, on remarque une tendance croissante du nombre de nuitées à partir de 2020 et la présence d'une saisonnalité d'une période d'un an (schéma répétitif) où le nombre de réservations atteint un pic chaque mois d'août.

Choix du modèle : Aux vues de la visualisation de la série temporelle, nous avons choisi un schéma additif. Dans un schéma additif, la série chronologique est modélisée comme la somme des composantes de la tendance, de la saisonnalité et d'erreurs (fluctuations aléatoires).

Création des modèles : Nous avons procédé à la décomposition de la série en isolant la tendance d'un côté et les coefficients saisonniers de l'autre. Pour ce faire, nous avons utilisé un modèle de régression linéaire pour nos deux composantes. Pour calculer la tendance, nous avons mis en place un modèle de régression linéaire par morceaux en utilisant la méthode des moindres carrés ordinaires. En ce qui concerne les saisons, nous avons soustrait la tendance estimée de notre modèle à la série temporelle initiale. En utilisant des variables indicatrices, représentant les différents moments saisonniers, nous avons pu obtenir les coefficients saisonniers. Ainsi, on observe sur la figure ci-contre une augmentation du nombre de nuitées au mois d'août par rapport à la moyenne (378 000 réservations en moyenne sur ce mois).

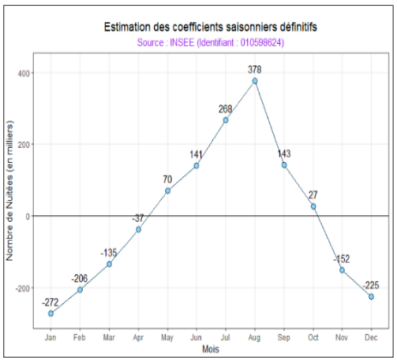

Figure 5 - Coefficients saisonniers

Vérification des modèles : Pour s'assurer de la validité du modèle, nous avons examiné les résidus pour repérer toute valeur mal ajustée et de détecter la présence d'informations non prises en compte par le modèle.

C. Conclusion

Ce projet m'a permis de prendre en main des méthodes statistiques plus complexes (décomposition de séries chronologiques) afin de réaliser des prédictions. De plus, lors de ce travail, j'ai dépassé la simple utilisation des méthodes statistiques en faisant un véritable effort de contextualisation des résultats.

SAÉ 3.04 – Conformité réglementaire pour analyser des données

A. Présentation

Ce module s'est présenté sous la forme d'enseignements et a fait l'objet d'une restitution de connaissances lors d'une séance d'examen sur feuille. Il avait pour objectif de nous sensibiliser aux enjeux de la réglementation générale sur la protection des données et aux problématiques de mise en conformité des données. Pour cela, nous avons échangé avec le Délégué à la Protection des données du Crédit Agricole de Normandie, venu spécialement pour l'occasion, afin d'échanger autour des défis et des bonnes pratiques en matière de protection des données personnelles.

B. Contenu de l'enseignement

Ces échanges avec le DPO du Crédit Agricole de Normandie ont eu deux objectifs : présenter la réglementation et partager les bonnes pratiques en matière de gestion des données.

La réglementation liée au traitement des données :

J'ai pu approfondir ma compréhension sur les missions de la Commission Nationale de l'Informatique et des Libertés (**CNIL**). En effet, la CNIL joue un rôle central dans la régulation et l'accompagnement du respect des données personnelles. Nous avons également introduit le Règlement Général sur la Protection des Données (**RGPD**), une législation européenne visant à encadrer le traitement des données au sein de l'Union européenne. Nous avons défini ce qu'est une donnée sensible, soulignant que, sauf exception, de telles données ne peuvent être traitées et analysées. Parallèlement, nous avons examiné les diverses bases légales sur lesquelles repose le traitement légitime de ces informations ainsi que l'ensemble des droits des individus, notamment le droit d'accès, à l'effacement ou encore à la limitation du traitement. En outre, nous avons évoqué les enjeux autour des transferts de données à l'étranger (par exemple, le **Data Privacy Framework**).

Les bonnes pratiques :

Nous avons également échangé autour des bonnes pratiques à mettre en place. Nous avons essayé de répondre à diverses questions : Comment traiter des données de façon responsable ? Quelles mesures de protection pouvons-nous mettre en œuvre ?

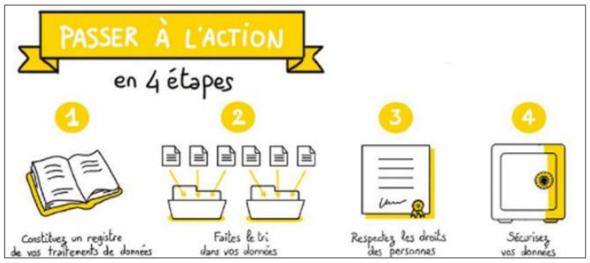

Figure 6 - Image réalisée par la CNIL montrant les 4 bonnes pratiques à suivre lors d'un projet

Pour garantir une gestion responsable des données, il existe différents points de vigilance : la durée de conservation des données, la sécurisation des informations, la mise en place de processus de tri dans les données ou encore la minimisation des données. Afin de respecter l'ensemble des réglementations durant un projet, j'ai retenu qu'il est nécessaire de débuter tout traitement par une phase de réflexion sur la licéité et la légitimité de la démarche (stratégie de **Privacy by Design**). En effet, l'analyse de données ne doit pas permettre d'identifier une personne spécifique, c'est pourquoi il est important de s'y pencher en début de projet au risque d'effectuer des traitements non conformes et donc non exploitables. À cette fin, il est conseillé de mettre en place un registre recensant l'ensemble des traitements exploitant des données (cf. Figure 6).

C. Conclusion

Ces échanges ont contribué à l'enrichissement de mon vocabulaire spécifique, un atout qui facilitera mon intégration dans le monde professionnel. De plus, cela a suscité une sensibilisation approfondie aux enjeux liés à la protection des données personnelles. Je suis désormais informé sur les ressources disponibles pour obtenir des conseils et des orientations en matière de protection des données.

SAÉ 4.01 – Expliquer ou prédire une variable quantitative à partir de plusieurs facteurs

A. Présentation

Au cours de ce projet, nous avons travaillé en groupes de trois étudiants à l'aide de **RStudio** sur l'analyse de données concernant des manchots. Les données provenaient du package « **palmerpenguins** » de **R**, qui recense des observations collectées par le Dr. Kristen Gorman sur l'archipel Palmer. L'objectif de ce projet était de créer des modèles de régression linéaire afin d'expliquer le poids d'un manchot à partir de différentes variables quantitatives et qualitatives. Nous avons développé différents modèles de diverses complexités, allant des modèles de régression linéaire simples aux modèles de régression linéaire de type ANOVA et ANCOVA.

B. Démarche et résolution du problème

Régression linéaire simple : Nous avons débuté par une analyse de la longueur de la nageoire, de la profondeur et de la longueur du bec, en évaluant leur association linéaire avec le poids des manchots. Après avoir appliqué des tests, notamment le test de Fisher, et comparé nos coefficients de détermination, nous avons conclu que la longueur de la nageoire était la variable la plus pertinente pour exprimer cette relation.

Analyse ANOVA : Nous avons ensuite procédé à une analyse de variance (ANOVA) pour étudier l'effet du sexe et de l'espèce sur le poids des manchots. Cette analyse nous a permis de déterminer que l'espèce était la variable ayant le plus d'influence sur le poids. Pour cette comparaison, nous avons utilisé la méthode du coefficient de détermination ajustée car la complexité de nos deux modèles était différente.

Analyse ANCOVA : Nous avons ensuite utilisé une analyse de covariance (ANCOVA) en intégrant la longueur de la nageoire, qui avait été sélectionnée précédemment, couplée à l'espèce ou au sexe. Cette analyse a évalué si l'inclusion de la longueur des nageoires, en association avec l'espèce ou le sexe, améliorait la capacité d'explication du poids des manchots. Nous avons utilisé des tests de Fisher pour comparer la significativité de nos modèles, tout en examinant les interactions significatives entre la longueur des nageoires et ces variables. La figure ci-contre met en évidence une interaction entre la longueur de la nageoire et l'espèce ; en effet, les droites de régression lissées ne sont pas parallèles.

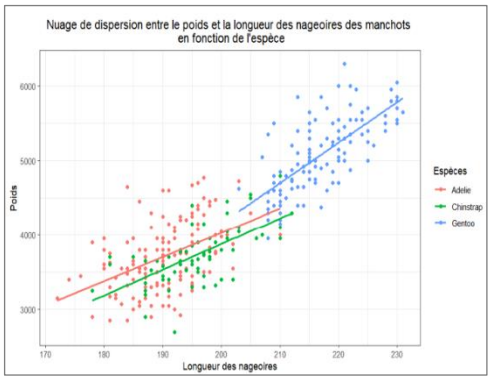

Figure 7 - Distribution entre le poids et la longueur des nageoires

Modélisation ascendante et descendante : Nous avons procédé à une sélection de modèles en utilisant une approche ascendante et descendante. Cette démarche a consisté à ajouter ou retirer sélectivement des variables explicatives du modèle de régression, en fonction de leur contribution à l'explication de la variation du poids des manchots. Cette méthode nous a permis de parvenir à un modèle final optimal, offrant un équilibre entre performance prédictive et complexité du modèle.

C. Conclusion

En résumé, ce projet m'a offert l'opportunité de maîtriser une gamme étendue d'outils statistiques, allant des différentes méthodes de comparaison de modèles (test de Student, test de Fisher, méthode AIC, etc.) aux techniques d'évaluation de la validité du modèle (normalité, homogénéité, gestion des valeurs atypiques). Travailler sur des modèles de diverses complexités m'a permis de mieux appréhender l'utilisation de ces méthodes.

SAÉ 4.02 – Reporting d'une analyse multivariée

A. Présentation

La cohorte **AGRICAN** est une cohorte prospective portant sur une population d'affiliés au régime agricole d'assurance maladie. Elle s'intéresse à la santé des agriculteurs et des secteurs connexes à l'agriculture. Pour ce projet, nous avons travaillé sur une sous-cohorte (près de 60 000 agriculteurs) de ce panel. Ces données portaient sur les activités agricoles pratiquées par les agriculteurs telles que l'élevage de bovins, de porcs, la culture de maïs, etc. L'objectif était de reconstruire la carrière des agriculteurs dans le but de regrouper les parcours agricoles similaires au moyen d'une méthode statistique de classification (méthode de **clustering des K-means**). Ce travail s'est effectué par groupe de deux étudiants sur le langage de programmation **R** et a donné lieu à un rapport d'analyse édité au moyen de **R Markdown**.

B. Démarche et résolution du problème

Notre première étape a été de prendre connaissance du questionnaire à partir duquel les réponses ont été recueillies. Nous avons pris soin de l'annoter afin de bien comprendre l'ensemble des variables disponibles dans le jeu de données. Ensuite, nous avons mis en place un script automatisé permettant de calculer la durée totale de la carrière de chaque agriculteur et la part de chaque activité sur ce temps de carrière (ratio d'activités). Cette étape a demandé un effort considérable en termes d'automatisation, nécessitant la mise en place de différentes fonctions pour traiter un ensemble important de variables.

Nous avons ensuite réalisé une analyse en composantes principales afin de mettre place un clustering en utilisant la méthode des **K-means** pour regrouper les agriculteurs en clusters (groupes d'individus) présentant des parcours agricoles similaires. L'algorithme des K-means itère successivement en ciblant des points de données comme centroïdes et en calculant leurs distances par rapport à tous les autres points. Ce processus se répète jusqu'à ce que la variance intra-cluster (somme des distances entre les points) ne diminue plus de manière significative à chaque itération, indiquant ainsi que les centroïdes sont stabilisés et que le nombre optimal de clusters est alors atteint. Enfin, nous avons examiné les caractéristiques communes des agriculteurs dans chaque cluster.

Le tableau (cf. Figure 8) ci-contre est un extrait des différents clusters obtenus. Dans la colonne « Moyenne », nous trouvons les ratios d'activité. Les ratios significativement inférieurs à la moyenne sont représentés en rouge, et ceux significativement supérieurs sont représentés en vert. Par exemple, pour l'élevage de bovins, la moyenne observée à travers tous les clusters était de 61,40 %. Cela signifie que, en moyenne, les agriculteurs de la cohorte consacrent 61,40 % de leur temps de carrière totale à l'élevage de bovins, que ce soit comme activité principale ou en conjonction avec d'autres pratiques agricoles durant cette période.

Table 4: Caractéristiques des clusters

	Moy c8	v-test8	Moy c6	v-test6	Moyenne
Bovins	9.25	-55.55	89.59	21.53	61.40
Moutons/chèvres	0.58	-11.53	5.74	-0.08	5.79
Cochons	1.26	-18.23	65.15	58.39	12.95
Chevaux	3.24	-13.00	61.13	59.00	11.14
Volailles	3.11	-15.78	58.38	45.74	14.07
Vigne	95.63	71.46	34.68	4.56	28.73
Mais	4.5	-31.84	25.92	-5.17	32.20
Bettraves	1.12	-19.55	63.71	49.81	14.87
Tournesol	0.55	-10.94	0.89	-7.15	4.45
Colza	0.34	-12.54	7.08	2.31	5.70
Tabac	0.78	-10.23	5.99	1.62	5.05
Arboriculture	2.16	-13.06	24.58	16.59	10.25
Prairies	14.39	-44.72	87.66	22.11	57.75
Autres.légumières	0.66	-11.50	10.62	6.28	6.31
Blé.ou.orge	7.27	-41.32	83.92	26.61	47.66
Cultures.sous.serres	0.27	-8.00	0.81	-4.60	3.01
Pois.fourragers	0.09	-9.84	1.36	-4.40	3.47
Pommes.de.terre	2.61	-17.77	70.59	52.63	15.86

Figure 8 – Extrait des clusters réalisés

C. Conclusion

En conclusion, ce projet a permis de reconstruire les carrières des agriculteurs et de les rassembler en groupes homogènes selon des parcours similaires. Il nous a fallu traiter une grande quantité de données, nécessitant une approche méthodique pour effectuer nos ratios. De plus, l'utilisation de la méthode de clustering des **K-means** a été une expérience enrichissante, me permettant de clarifier les mécanismes sous-jacents du clustering en passant par les étapes de sélection des composantes, du choix du nombre de clusters optimal et d'analyse.